FACULTY OF ENGINEERING AND ARCHITECTURE

IDLAB – INTERNET TECHNOLOGY AND DATA SCIENCE LAB - IMEC

Aranka Steyaert, Pieter Audenaert and Jan Fostier

ACCURATE DETERMINATION OF NODE AND ARC MULTIPLICITIES IN DE BRUIJN GRAPHS USING CONDITIONAL RANDOM FIELDS

Background

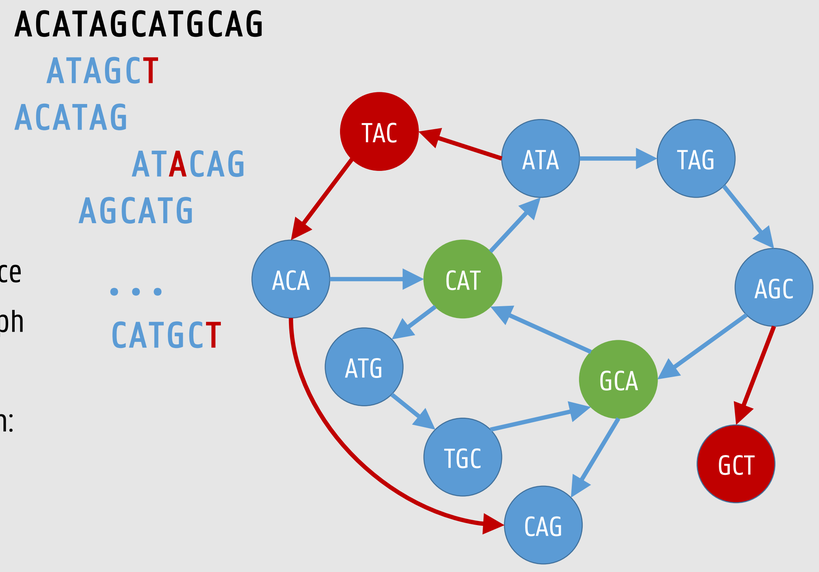
Genome sequencing

- Produces millions of short reads (50-250 bp length).
- Origin in genome sequence unknown.
  → "billion pieces genomic puzzle"
- Extra difficulty: reads contain errors (1-2%)
- Determine overlap between reads
  → all to all comparison inefficient
  → obtain 'k-mers' from all reads: sub-sequences of equal length k
  count occurrence and occurrence of k-1 overlap

De Bruijn Graph

Representation of the k-mers and their overlap

- **nodes**: k-mers
- **arcs**: overlap of k-1
- **read support**: occurrence of k-mer/overlap in read set
- **multiplicity**: occurrence of k-mer/overlap in original sequence
  - original sequence present as walk through graph
- **conservation of flow of multiplicity**:
  if the full original genome is represented by de Bruijn graph:
  NODE MULTIPLICITY = ∑ INCOMING ARC MULTIPLICITIES
  = ∑ OUTGOING ARC MULTIPLICITIES



K-mer spectrum

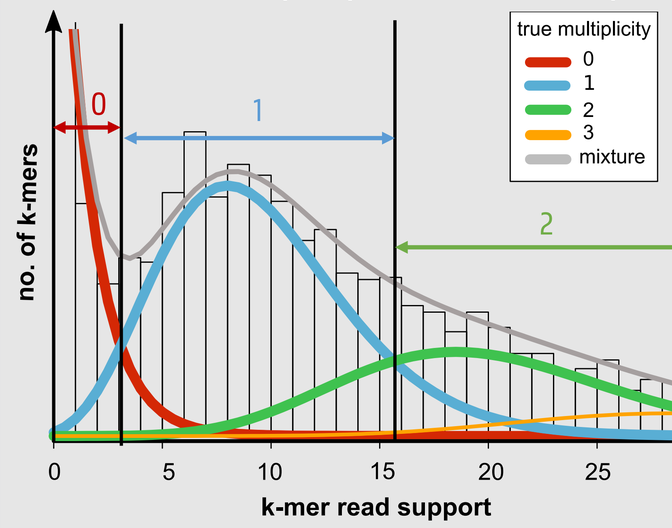
Histogram of read support of all k-mers

Fit mixture model to these counts

→ One distribution per multiplicity

→ Determine cut-off values and create intervals of multiplicity

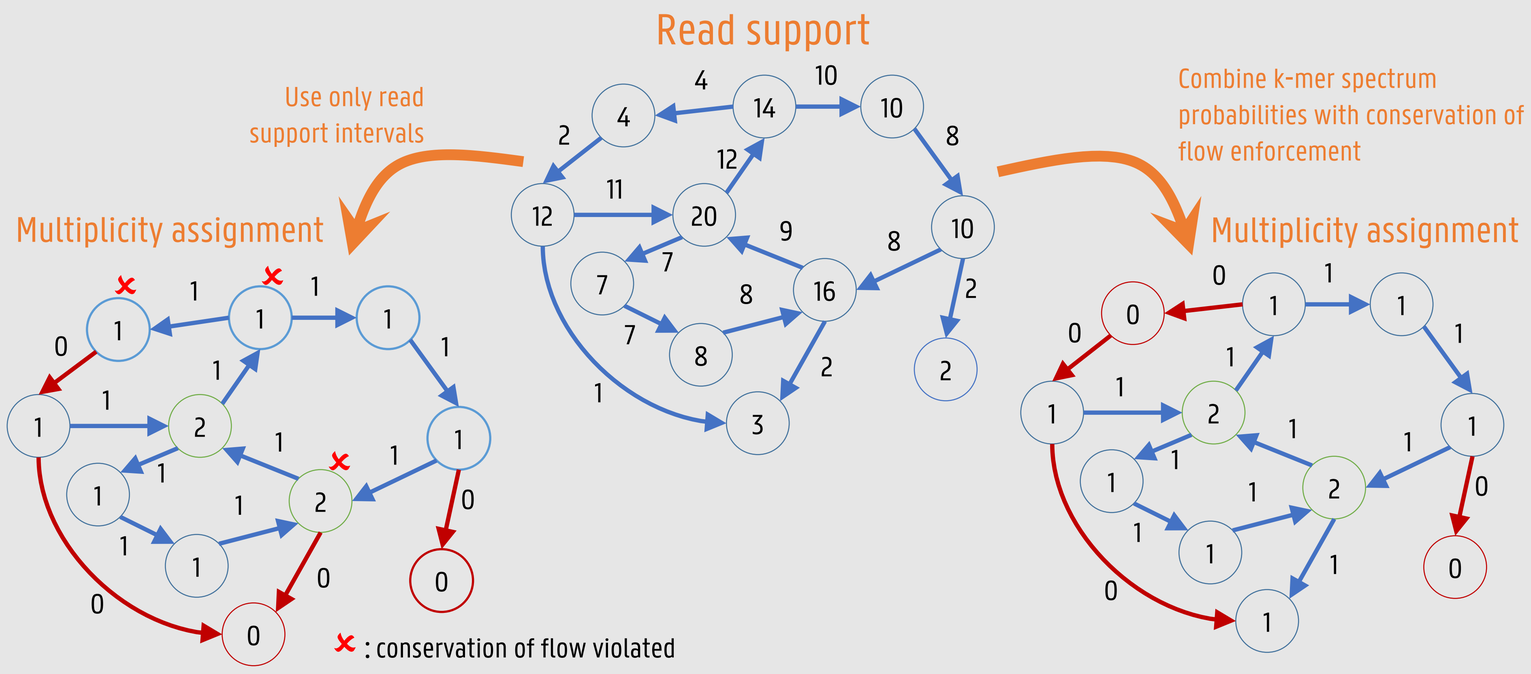
Issue: different multiplicity distributions overlap



Motivation

State of the art methods assign multiplicities based on cut-off values in k-mer spectrum

Incorporating the conservation of flow of multiplicity property enforces more correct multiplicity assignments



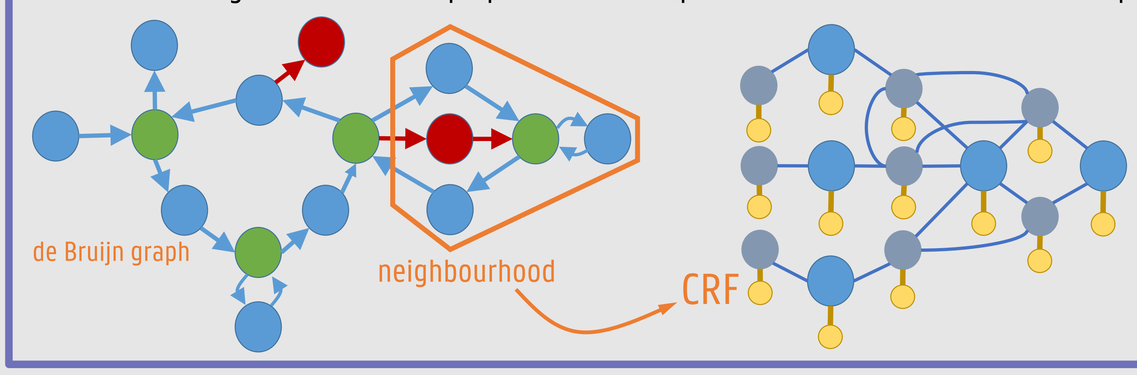
Results

Fraction of correctly assigned multiplicities in the de Bruijn graph (Illumina data, 30x coverage)

	K-mer spectrum	+CRF
S. enterica	89%	95%
P. aeruginosa	80%	98%
B. dentium	81%	97%
E. coli	97%	98%
C. elegans	55%	63%
H. sapiens chr. 21	69%	77%
D. melanogaster	67%	73%

Conditional Random Fields

Probabilistic framework that allows us to combine k-mer spectrum based probabilities with probabilities on neighbourhoods of nodes that enforce the conservation of flow. CRFs are a proved technique in image segmentation to incorporate information embedded in neighbourhoods of superpixels. A whole spectrum of efficient inference techniques has already been developed.



multiplicities $\mathbf{Y} = \{ Y \mid \text{all nodes and arcs} \}$

read support $\mathbf{X} = \{ X \mid \text{all nodes and arcs} \}$

$\varphi(\mathbf{Y}|\mathbf{X})$

$\varphi(Y_{\text{node}}, \mathbf{Y}_{\text{in-arcs}})$

$\varphi(Y_{\text{node}}, \mathbf{Y}_{\text{out-arcs}})$

$$P(\mathbf{Y}|\mathbf{X}) = \frac{1}{Z(\mathbf{X})} \prod_i \varphi_i(\mathbf{D}_i)$$

Contact

aranka.steyaert@ugent.be
www.idlab.ugent.be, www.imec-int.com

Universiteit Gent

@ugent @IDLabResearch

Ghent University

GHENT UNIVERSITY imec IDLab INTERNET & DATA LAB